Bibliographic Reference Segmentation for Bibliometrics

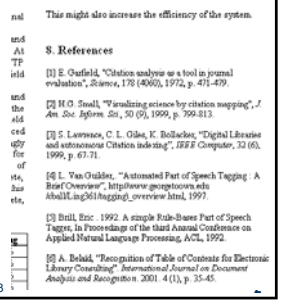
Abdel Belaïd, Dominique Besagni, Nelly Benet

Bibliographic Reference Segmentation for Bibliometrics

A.Belaïd, D. Besagni, N. Benet
LORIA - INIST

Bibliographic references

- Definition
  - Citations mentioned at the end of scientific publications
  - One of the structural elements of a standard scientific article that can be used for analysis


This might also increase the efficiency of the system.

8. References

[1] E. Garfield, "Citation analysis as a tool in journal evaluation", *Science*, 178 (4060), 1972, p. 471-479.

[2] M.O. Small, "Visualizing science by citation mapping", *J. Am. Soc. Inform. Sci.*, 50 (9), 1999, p. 799-813.

[3] S. Lawrence, C. L. Giles, K. Bollacker, "Digital Libraries and autonomous Citation indexing", *IEEE Computer*, 32 (6), 1999, p. 67-71.

[4] L. Van Guilder, "Automated Part of Speech Tagging : A Brief Overview", http://www.georgetown.edu /cball/ling361/tagging_overview.html, 1997.

[5] Brill, Eric . 1992. A simple Rule-Based Part of Speech Tagger, in Proceedings of the third Anaual Conference on Applied Natural Language Processing, ACL, 1992.

[6] A. Belaïd, "Recognition of Table of Contents for Electronic Library Consulting", *International Journal on Document Analysis and Recognition*. 2001. 4 (1), p. 35-45.

Citation vs Reference

- Its a difference of perspective between the author cited and the researcher citing the author
  - For the author citing :
    - It is a reference to the author cited
  - Inversely, for the author cited
    - it is a citation by the citing author
- The INIST (Institute for Scientific and Technical Information)
  - Engaged a research on these references because of the interest of the citations in the bibliometrics and scientometrics studies

Bibliometrics

- Before 1960
  - Defined as quantitative research of every thing concerning the science and for which we can attach numbers
  - Used on PASCAL(INIST), MEDLINE(US), INSPEC(UK)
- Defined by Pritchard, 1969 as
  - The application of mathematics and statistical methods to :
    - books, articles and other communication means
  - It can be particularly applied to :
    - the study of the publishing of scientific papers considered as an indicator of the scientific activity

Bibliometrics

- The creation on 1960 by Eugene Garfield of the foundation of the Institute for Scientific Information (ISI)
  - Will give a new measure element : the citation
- Used at first :
  - Only as a tool for information retrieval
  - The citation has become an important criterion
    - because it allows to distinguish among different publications those which received the approbation of the scientific community
    - also used to appraise scientific journals especially with the impact factor calculated as the average number of citations a paper receives over a period of 2 years

Scientometrics
Two approaches

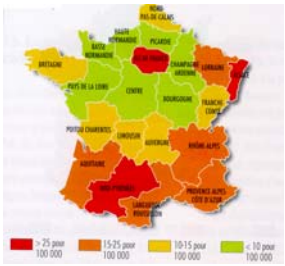
to measure the impact of an institution in one domain

1 Bibliometrics
- Surface Info : distributions, frequencies per date, journal, author, domain
  - Lab Production, Country

2 Infometrics
- Info on the content : to highlight themes, co-citations
  - Relational Indicators between themes and authors

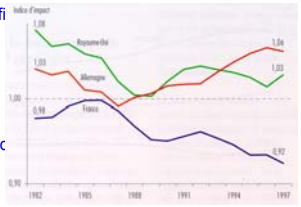
Bibliometrics
Example

- Nb of thesis / population
- Predominance of the Ile de France (Paris)
- Effects on the neighbor regions



Bibliometrics
Example

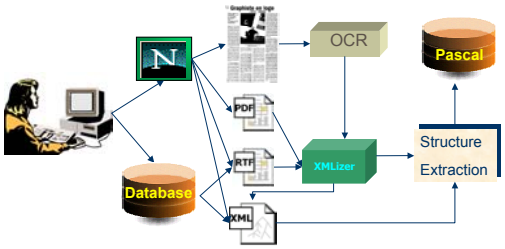
- Impact Index of the scientifi publications in the E.U.
- Impact = average nb of references per article
- One remarks a continuous decreasing in France, i.e. French articles are less anc less read and referenced
- Several possible causes...



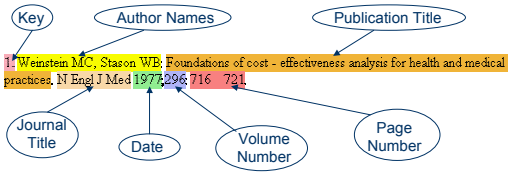
Bibliometrics at the INIST
Institute for Scientific and Technical Information

- Has
  - A basis covering the most part of the scientific and technical research in the world
  - Two multilingual and multi-disciplinary data bases :
    - PASCAL and FRANCIS
- Aim
  - Automatic updating of these databses
  - To bring additional value to the bases by :
    - Bibliometrics : idea on the best themes and famous authors

Reference data processing
The global architecture



Structure Recognition
The basic idea



Data and Method

- The raw data, obtained by OCR :
  - Set of "well-formed" XML documents in which each reference from the same article is singled out
- Character set used : ISO-latin 1 (standard ISO 8859-1).

```
<INFCOM fic="1998/refm278.dat">
<NUMACQ><CLEA>35400007110423</CLEA><CLEB>0030</CLEB></NUMACQ>
<REFBIB copie="0" >1 American Cancer Society. Cancer Facts and Figures-1997, American Cancer Society: Atlanta, 1997.</REFBIB>
```

Problems of several kinds
Due to the digitisation

- Confusions :
  - I→L, C →(, 8 →B…

Ami J, Fredricson Overo K, Hyllel J. Olsen R (1984) Changes In rai dopamine and serolonin funcllon In vivo after prolonged adminisirallon oflhe spécifie 5 - HT uplake Inhibllor, cllalopram Psychopharmacology 84:457 - 4Ê5

Balfour DJ. Graham ÇA. Vale AL (198B) Sludies on Ihe possible rôle ofbrain 5 - HT syslems and adrenocorllcal acilvlly In behavloural responses lo nicolìne and dlazepam In an elevaled X- maze. Psychopharmacology 90:528 - 532

Balldin J, Berggren U. Engel J. Enksson M. Hard E. Soderpalm B (1994) Effecl of cllalopram on alcohol Inlake In heavy drinkers. Alcohol Clin Exp Res 18:1133 - 113B

Problems of several kinds
Due to the heterogeneity of the data

- Even though the model of the citation depends on the journal in which it is published
  - The structure of a reference may vary greatly from one paper to another in the same journal

EVANS, W.H. Intercellular Communication. (1997). The Roles and Structure of Gap Junctions. In: Principles of MedicalBiology, JAI Press Inc., 609 - 628.

Wilson JD: Androgens. In, Hardman JG, Limbird LE (eds.): Goodman and Gilman's The Pharmacological Basis of Therapeutics. Ninth edition. New York: McGraw Hill, 1996;1441 - 1457.

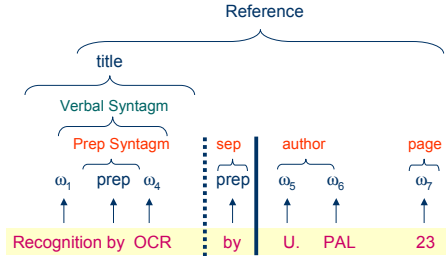
SH. Harper J Chem Soc (1939) 1099.

Still, there are a few regularities
Within the same paper, references have:

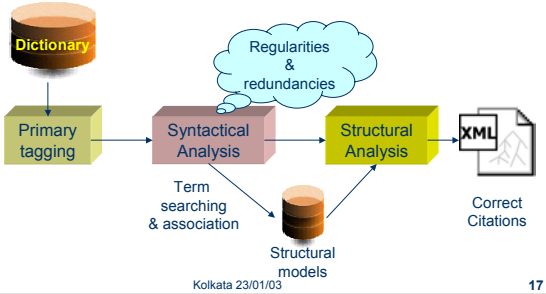
- Same structure: for the same type of quoted document
- Authors' names: in the beginning
- Date: very limited number of positions: after the authors, after the journal title, after the pagination
- Page number: at the end

1. Weinstein MC, Stason WB: Foundations of cost - effectiveness analysis for health and medical practices. N Engl J Med 1977;296: 716 - 721.
2. Russell LB. Gold MR, Siegel JE, Daniels N, Weinstein MC: The role of cost - effectiveness analysis in health and medicine. JAMA 1996;276:1172 - 1177,
3. Alexander B, Nasrallah HA, Perry PJ, Liskow BI, Dunner FJ: The impact of psychopharmacology education on prescribing practices. Hosp Community Psychiatry 1983;34:1150 - 1153.

The approach
Part of Speech Tagging



The approach
Morpho-syntactical Labeling



The approach
Primary Labeling

Tag	Meaning	Tag	Meaning
AN	Alphanumeric string	IN	Expression "In:"
CC	Connector (and, & …)	IT	Initial
CWC	Common noun, capitalised	JM	Journal marker
CWL	Common noun, lowercase	NM*n*	Number (*n* digits)
CWU	Common noun, uppercase	PN	Proper name
EA	Expression "*et al.*"	PU*s*	Punctuation mark *s*
ED	Editor (Ed., Eds.)	UN	Unknown

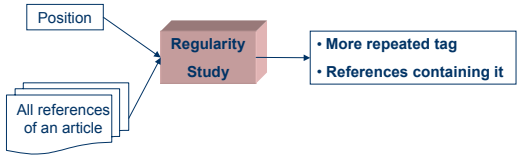
The approach
Primary Labeling

1. Weinstein MC, Stason WB: Foundations of cost - effectiveness analysis for health and medical practices. N Engl J Med 1977;296: 716 - 721.

1/**NM1** ./**PU.** Weinstein/**PN** MC/**PN/IT** ,/**PU,** Stason/**PN** WB/**IT** :/**PU:** Foundations/**CWC** of/**CWL/PR** cost/**CWL** -/**PU-** effectiveness/**CWL** analysis/**CWL** for/**CWL/PR** health/**CWL** and/**CC** medical/**CWL** practices/**CWL** ./**PU.** N/**IT** Engl/**PN/JM** J/**JM/IT** Med/**PN/JM** 1977/**NM4** ;/**PU;** 296/**NM3** :/**PU:** 716/**NM3** -/**PU-** 721/**NM3** ./**PU.**

Syntactical analysis

1. Regularity principle
   - Numerical fields
   - Alpha-numerical fields
   - Examples: keys, dates



Syntactical analysis

2. Syntactical rules (Author)
   - Reduction
     PN!IT + PU- + PN!IT => PN
   - Formation
     IT + PN => AU
   - Agglomeration
     AU + PU[.,] + AU => AU

<CLE> 1/NM1 ./PU. **</CLE>** **<AU>** Weinstein/PN MC/PN/IT ,/PU, Stason/PN WB/IT **</AU>** :/PU: **<TIP>** Foundations/CWC of/CWL/PR cost/CWL -/PU- effectiveness/CWL analysis/CWL for/CWL/PR health/CWL and/CC medical/CWL practices/CWL **</TIP>** ./PU. **<JN>** N/IT Engl/PN/JM J/JM/IT Med/PN/JM **</JN>** **<DA>** 1977/NM4 **</DA>** ;/PU; **<VOL>** 296/NM3 **</VOL>** :/PU: **<PG>** 716/NM3 -/PU- 721/NM3 **</PG>** ./PU.

Structural Analysis

- Syntactic analysis: limits:
  - Some terms are unknown
  - The title has a too complex structure
  - Confusion between the publication year and the pagination
- Proposed two kind of models:
  - Inter-fields : searching for limits
  - Intra-fields : confirmation of these limits

Inter field modeling

- We used a pair modeling revealing the association of consecutive fields. This gives the sequence of possible consecutive fields with their separators

Tag 1	Tag 2	Percentage	Position	Simple separator	Double separator
AU	DA	97.83	1.00	(93	)(2,(2
DA	TIP	56.52	1.96		) 56
DA	AU	2.17	2.00		) 2
TIP	JN	56.52	2.62	56	
AU	TIP	2.17	3.00	2	
JN	VOL	65.22	3.20	,60	
VOL	PG	95.65	3.89	,93 ,2	

3. Building model
2. Selecting separators
1. Suppressing invalid couples

Inter field correction

Cohen C, Perrault G Sanger DJ (1998) Preferential involvement of D3 versus D2 dopamine receptors in the effects of dopamine receptor ligands on oral ethanol self - administration in rats. Psychopharmacology 140 478 - 485

Cohen C, Perrault G Sanger DJ (1998) Preferential involvement of D3 versus D2 dopamine receptors in the effects of dopamine receptor ligands on oral ethanol self - administration in rats. Psychopharmacology 140 478 - 485

Intra field modeling

- Once obtained the field identity and limits, we try to find out the kind of elements used in the field and their structure in terms of sequence and separators

First author	
Pattern:	PN PU, IT PU.
Separator:	PU,
Next authors	
Pattern:	PN PU, IT PU.
Separator:	PU,
Last author	
Pattern:	PN PU, IT PU.
Connector:	and

Intra field correction

23. Kozaesky, K. F., and Wilson, J. M. Gene therapy: adenovirus vectors. Curr. Opin. Genet. Dev., 3: 499 - 503, 1993.

23. Kozaesky, K. F., and Wilson, J. M. Gene therapy: adenovirus vectors. Curr. Opin. Genet. Dev., 3: 499 - 503, 1993.

Experiments and results

- 140 journals of pharmacology
  - 64 articles chosen at random from the original set. It contains 2,575 references

Fields	Complete	Partial	Not found	Wrong
Authors	90.2%	6.6%	0.3%	2.9%
Title	82.4%	15.4%	1.7%	0.4%
Journal	92.4%	2.9%	3.2%	1.5%
Date	97.7%	0.0%	2.3%	0.0%
Volume	93.6%	0.4%	5.8%	0.2%
Pagination	94.7%	0.6%	4.3%	0.4%
Whole Reference	75.9%	18.8%	0.0%	5.3%